Разведанализ данных

В исследовании мы используем усредненные данные по **50 штатам США с 2010 по 2019 год.**

Чтобы изучить экономические условия штата, мы включили такие характеристики как количество совершенных насильственных преступлений, уровень бедности и безработицы, а также доход и реальный ВВП населения. Кроме того, добавили несколько контрольных переменных: плотность населения, число иммигрантов в штате, количество браков и разводов, а также среднюю температуру в штате.

Данные были собраны с разных ресурсов, так что требовали некоторых преобразований для дальнейшего анализа.

Предварительный просмотр данных:

Show 10 **entries** **Search:**

	State	Vcrime	PR	UR	PIPC	RGDP	Density	Immigrants	Marriage	Divorce
1	Alabama	462.25	0.1805	0.0657017690565541	38436.5	193018.02	96.8	3908	7.54293619722883	11.3
2	Alaska	702.87	0.1084	0.0676240971098351	55765.2	54741.09	1.25	1568	7.24240670261611	11.5
3	Arizona	432.71	0.1673	0.0688650774555567	39264.4	286210	59.6	18456.6	5.69068450216861	10.0
4	Arkansas	506.21	0.1813	0.0558507627847626	38907.7	111633.63	56.95	2906.4	9.85616031205131	11.7
5	California	433.69	0.15	0.0764721540817871	53327.5	2335825.91	246.4	204678.6	6.11870791926213	7.7
6	Colorado	336.26	0.1173	0.0527386225249409	50676.4	306193.02	52.1	13022.5	7.06243878718641	8.6

Showing 1 to 6 of 6 entries Previous 1 Next

Рассмотрим обощающие данные по всем переменным:

Переменная State

```
   Length     Class      Mode
       50 character character
```

Текстовая переменная, обозначающая название штата.

```
 [1] "Alabama"        "Alaska"         "Arizona"        "Arkansas"
 [5] "California"     "Colorado"       "Connecticut"    "Delaware"
 [9] "Florida"        "Georgia"        "Hawaii"         "Idaho"
[13] "Illinois"       "Indiana"        "Iowa"           "Kansas"
[17] "Kentucky"       "Louisiana"      "Maine"          "Maryland"
[21] "Massachusetts"  "Michigan"       "Minnesota"      "Mississippi"
[25] "Missouri"       "Montana"        "Nebraska"       "Nevada"
[29] "New Hampshire"  "New Jersey"     "New Mexico"     "New York"
[33] "North Carolina" "North Dakota"   "Ohio"           "Oklahoma"
[37] "Oregon"         "Pennsylvania"   "Rhode Island"   "South Carolina"
[41] "South Dakota"   "Tennessee"      "Texas"          "Utah"
[45] "Vermont"        "Virginia"       "Washington"     "West Virginia"
[49] "Wisconsin"      "Wyoming"
```

Имеет 50 уникальных значений.

Переменная Vcrime

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  121.7   261.1   337.0   362.9   460.5   702.9
```

Числовая переменная, указывающая количество совершенных насильственных преступлений (violent crimes) на 100 тысяч населения.

Значения колеблются в диапазоне от 122 до 703, где среднее (363) и медианное (337) близки друг к другу.

На графике видно, что количество регистрируемых наблюдений в целом находится на достаточно высоком уровне, что еще раз подтверждает выводы СМИ о существующей проблеме с насильственными преступлениями на территории Соединенных Штатов Америки. Статистические службы также зафиксировали и несколько экстремально высоких значений за все время наблюдения (более 600 преступлений). Подобные выбросы можно объяснить снижением затрат на финансирование правоохранительных органов, усилением расслоения общества, экономическими шоками. Следует заметить, что это могло произойти и вследствие ошибок при фиксации и регистрации насильственных преступлений.

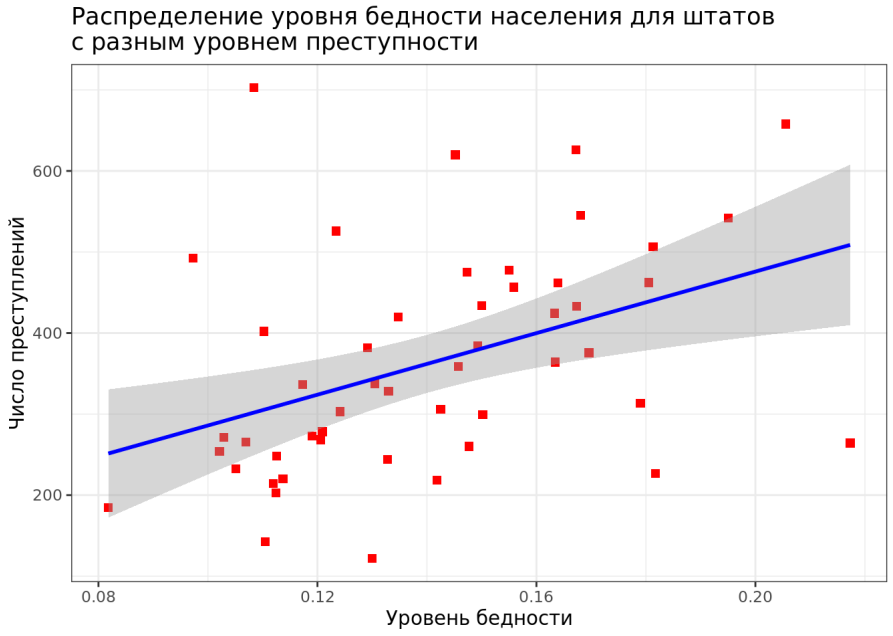
Переменная PR

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.0818  0.1146  0.1383  0.1405  0.1634  0.2173 
```

Poverty rate - числовая переменная, обозначающая уровень бедности в долях, то есть от 0 до 1.

В США усредненное за 10 лет минимальное значение 0.0818, а максимальное - 0.2173.

На графике рассеивания мы можем увидеть распределение уровня бедности относительно уровня преступности:



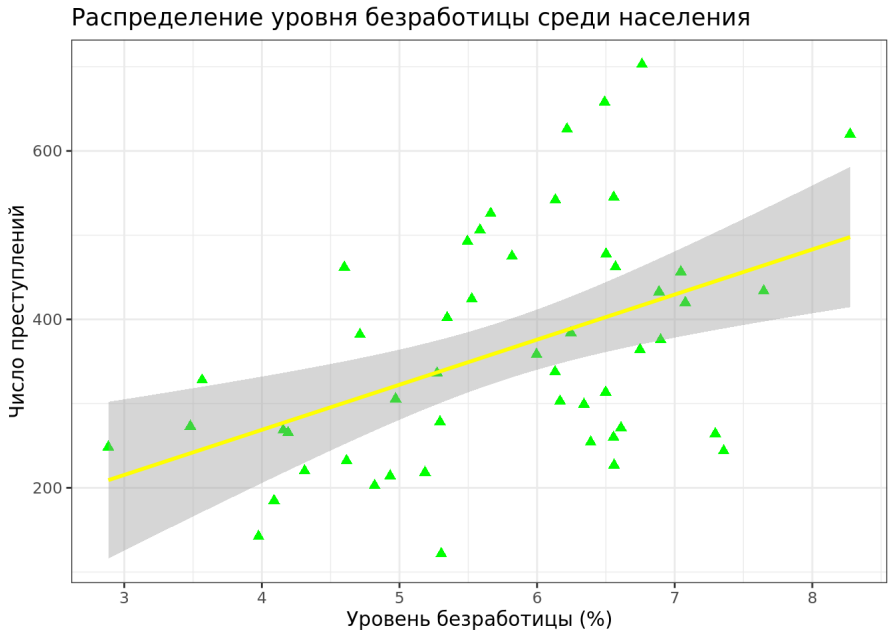
В целом, почти все штаты показывают неплохие - относительно мировых - результаты по уровню бедности. Тем не менее, есть штаты, которые за десять лет наблюдений все так же имеют достаточно высокую (близко к 0,2 и выше) долю населения, находящегося за чертой бедности. Это может объясняться локальными экономическими, географическими и климатическими особенностями, которые мешают развитию штата.

Переменная UR

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.02884 0.04942 0.06064 0.05755 0.06559 0.08275 
```

Unemployment rate - уровень безработицы в штатах в долях. Значения колеблются от 2.9% до 8.3%.

Точечный график показывает слабую корреляцию с зависимой переменной:



Следует заметить, что уровень безработицы в США в среднем держится на оптимальном уровне, потому что падения ниже 5%, как правило, приводят к снижению общего уровня выпуска и усилению инфляции. Подобные значения также говорят об эффективности работы Федеральной Резервной Системы.

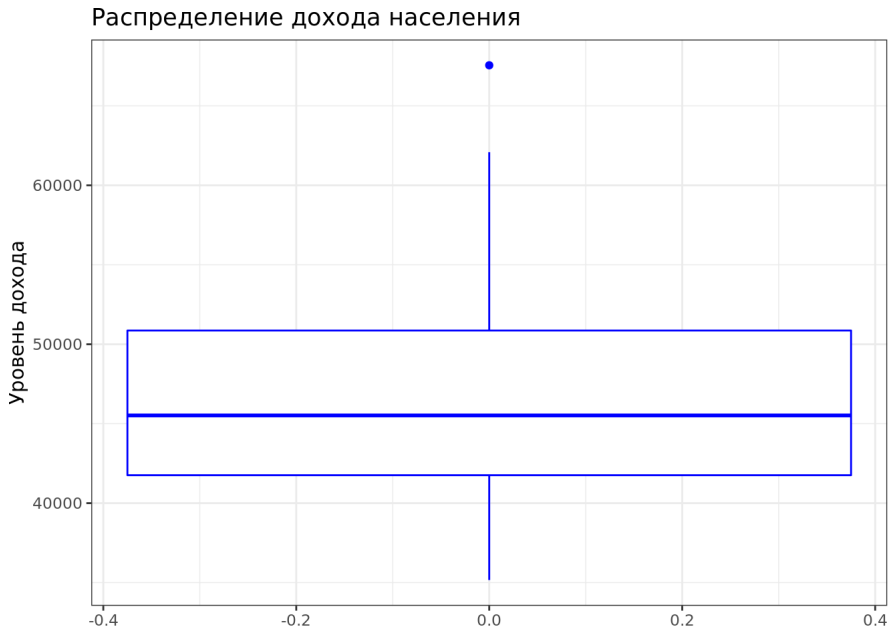
Переменная PIPC

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  35163   41759   45521   46785   50864   67553
```

Personal income per capita - числовое значение, выражающее доход на душу населения в долларах.

Максимальное значение (67553) почти в два раза больше минимального (35163), хотя среднее и медианное значения - чуть больше 45000. То есть значений, близких к максимальному, довольно мало.

Рассмотрим график типа boxplot:



Уровень дохода приходится в основном на промежуток от 43 до 51 тысячи долларов на душу населения. Это позволяет сделать вывод о том, что доходы жителей большинства штатов находятся именно в этом отрезке.

Тем не менее присутствует значение, которое выделяется из массы остальных. Проверим штаты с уровнем дохода больше 60000:

Show 10 **entries** **Search:**

	State	PIPC
1	Connecticut	67553.1
2	Massachusetts	62083.9

Showing 1 to 2 of 2 entries

Previous 1 Next

Выбросы в данных - значение уровня дохода в двух штатах.Большую долю в экономике Коннектикута составляет доход от военно-промышленных корпораций, так как это главный арсенал страны, а также страховой бизнес - один из наиболее популярных в штате. Массачусетс - один из мировых центров биотехнологий, искусственного интеллекта и венчурного капитала.

Выбросы можно объяснить спецификой показателя дохода: он фиксирует все доходы человека (зарплата, доходы от аренды, дивиденды, трансферты и другие виды). В данном случае это влияет на корреляцию между переменными, так что попробуем прологарифмировать значение PIPC. Логарифм делает показатели относительными, что позволяет избавиться от выбросов.

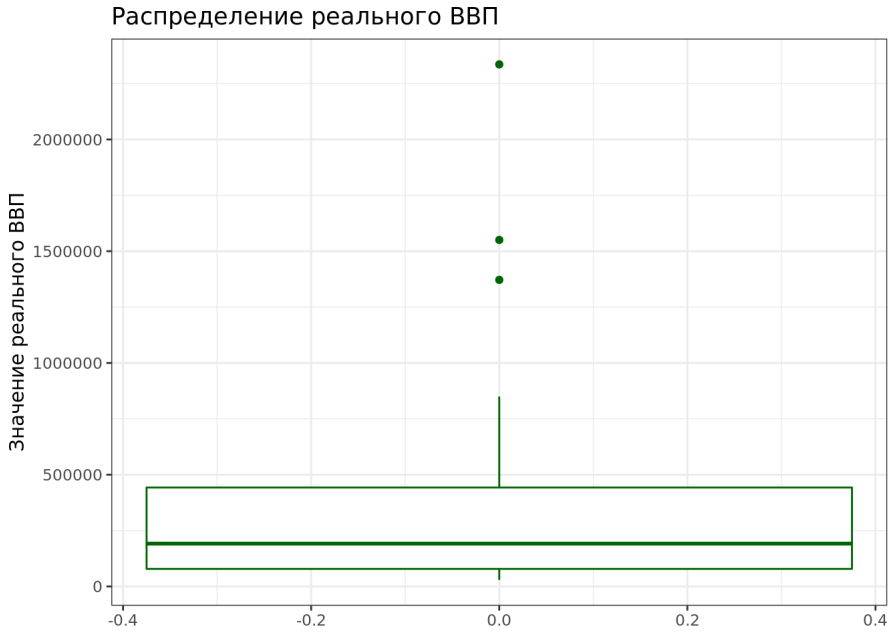
В рамках показателя Personal Income per Capita наблюдается положительная тенденция: усреднённые за 10 лет данные показывают, что доход на душу населения находится на достаточно высоком уровне почти во все штатах, что свидетельствует о приемлемом состоянии экономики. Кроме того, следует отметить, что подобные показатели достигаются, в том числе, благодаря западной контрактной системе с фиксированными почасовыми ставками.

Переменная RGDP

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  29180   78561  191783  339150  442790 2335826 
```

Real GDP - числовое значение реального ВВП штата. За 10 лет он колеблется от 29180 до 2335826.

Максимальное значение = 2335826, а значение 3го квартиля = 442790. Это говорит об очень сильных выбросах. Посмотрим график boxplot:



Тут видны выбросы в лице предположительно 3 штатов. Проверим значения выше 700000:

Show 10 **entries** **Search:**

	State	RGDP
1	California	2335825.91
2	Texas	1550373.9
3	New York	1371720.34
4	Florida	849951.18
5	Illinois	741158.3

Showing 1 to 5 of 5 entries Previous 1 Next

Действительно, мы видим, что 3 штата - Калифорния, Техас и Нью-Йорк - отличаются от других штатов тем, что средний реальный ВВП с 2010 по 2019 год был выше 1300000, а у остальных штатов значительно ниже.

Нетрудно объяснить подобное отклонение реального ВВП:

- Калифорния - крупнейшая экономика в рамках Соединенных Штатов Америки, пятая по показателям ВВП даже среди стран. Одними из драйверов подобного успеха являются численность населения и диверсификация экономики (финансовый сектор и недвижимость, информационная отрасль, производственный сектор вносят наибольший вклад).
- Экономика Техаса - вторая по размерам после калифорнийской. Данный штат стремительно развивается уже многие годы, и это однозначно отражается на его показателях: первое место по темпам роста ВВП, первое место по экспорту в США. Экономика Техаса во многом зависит от торговли и обилия крупного бизнеса.

- Стоит отметить и экономику штата Нью-Йорк - третью по размерам в США и двенадцатую - по всему миру. Штат является центром всех финансовых операций: подтверждением тому служит Нью-Йоркская фондовая биржа. Кроме того, Нью-Йорк - крупный морской порт.

Однако эти выбросы могут отрицательно повлиять на наши будущие оценки, поэтому логарифмируем переменную.

В рамках Real GDP наблюдается тенденция к росту показателя в зависимости от размера и численности населения штата, его ранга бизнес-среды, скопления высокотехнологичных производств и размерности рынка венчурного капитала.

Переменная Density

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.25   45.76  105.15  200.74  217.05 1229.25 
```

Density - плотность населения на единицу площади в квадратный милях.

Разброс данных велик: от 1.25 до 1229.25. При этом третий квантиль составляет всего 217.05, скорее всего присутствуют некие выбросы. Проверим значения в два раза больше границы квантиля:

Show 10 **entries** **Search:**

	State	Density
1	New Jersey	1229.25
2	Rhode Island	1039.75
3	Massachusetts	870.3
4	Connecticut	741.4
5	Maryland	615.45

Showing 1 to 5 of 5 entries Previous 1 Next

Мы видим 5 штатов с высоким уровнем плотности. Несложно объяснить подобные выбросы: представленные в списке штаты являются одними из самых маленьких по площади в США, но не самыми малочисленными по населению.

Данные нарастают относительно пропорционально и имеют логические объяснения, поэтому оставим их в изначальном виде.

Переменная Immigrants

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
   449.9   3124.7   6432.0  21081.0  18599.3 204678.6 
```

Immigrants - количество иммигрантов в штате, которые получили законный статус иммигранта/находятся в США на временной основе (студенты, работники, туристы)/получили одобрение на убежище/были натурализованы.

Снова сталкиваемся с проблемой разброса данных: максимальное значение в 10 больше среднего. Оценим значения, большие хотя бы в 2 раза:

Show 10 **entries** **Search:**

	State	Immigrants
1	California	204678.6
2	New York	140910.4
3	Florida	116317.5
4	Texas	99873.1
5	New Jersey	53156.2

Showing 1 to 5 of 5 entries Previous 1 Next

Выделяются 5 довольно крупных и популярных штата. Объяснить выбросы в количестве иммигрантов не составляет трудностей: большая часть представленных в списке штатов являются наиболее развитыми и привлекательными для миграции из-за высокого уровня дохода и лучших условий жизни. Эти штаты являются домом для крупнейших стартапов, имеют лучшую систему образования и идеальные карьерные перспективы.

Данные выбросы объяснимы и логичны, поэтому оставим их в изначальном виде.

Переменная Marriage

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  5.260   6.070   6.626   7.360   7.209  31.503 
```

Marriage - количество браков на 1000 человек.

При максимальном значении 31.5 граница третьего квантиля находится на допольно низком уровне (7.4), возможны выбросы. Проверим значения, выбивающиеся больше чем в два раза:

Show 10 entries Search:

	State	Marriage
1	Nevada	31.502662549274
2	Hawaii	16.2975960934697

Showing 1 to 2 of 2 entries Previous 1 Next

Видим два штата с большой разнице в даннах, однако и это довольно легко объяснить. Невада и Гавайи уже длительное время являются лидерами по количеству браков. Лас-Вегас (штат Невада) славится самым простым законодательством по отношению к заключению брака - это штат “быстрых” браков. Гавайи тоже появились в этом списке неслучайно: штат является экзотическим направлением для многих пар, желающих заключить брак в необычных условиях и при различных сюжетах на любой бюджет.

Оставим данные в исходном формате.

Переменная Divorce

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  6.150   7.825   9.100   9.055  10.088  12.250 
```

Divorce - количество разводов на 1000 человек.

Нет никаких выбросов или отличительных свойств. Данные предположительно нормально распределены, что не мешает в дальнейшей модели.

Переменная Temperature

Temperature - средняя температура в штате по Фаренгейту.

Чтобы было легче вопринимать, переведем данные в градусы по Цельсию и посмотрим основную информацию.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -1.633   8.068  11.333  11.655  15.194  22.183 
```

В целом, мы видим, что нет никаких выбросов и значения более или менее нормально распределены, оставляем значения в исходном виде - температура по Фаренгейту.

Корреляция между переменными

Для подробного анализа мы также проверили корреляцию между зависимой переменной Vcrime и объясняющими и контрольными переменными:

```
             Vcrime       PR       UR     PIPC     RGDP Temperature  logPIPC
PR           0.41702                                                       
UR           0.46002  0.51298                                              
PIPC        -0.17195 -0.74873 -0.19432                                     
RGDP         0.14664  0.09647  0.33579  0.25378                           
Temperature  0.30612  0.55985  0.30582 -0.38699  0.26231                  
logPIPC     -0.17469 -0.77363 -0.22181  0.99609  0.26278    -0.39801      
logRGDP      0.22073  0.13365  0.42050  0.17917  0.82532     0.37939  0.18031
```

```
                RGDP  Density Immigrants Marriage  Divorce
Density      0.19863                                      
Immigrants   0.94780  0.25143                             
Marriage    -0.14353 -0.18209   -0.08215                  
Divorce     -0.25462 -0.46456   -0.26433  0.23890         
Temperature  0.26231  0.11879    0.25057  0.07890  0.17616
```

Коэффициент, близкий или больший по модулю 50, означает высокую взаимосвязь между переменными. Ниже 50, соответственно, свидетельствует о слабой взаимосвязи.

На данной схеме мы можем увидеть, что существует довольно высокая корреляция переменной Vcrime с переменными PR (0,42) и UR (0,46). С переменными Temperature (0,31) и Divorce (0,29) прослеживается средняя корреляция. Остальные доволно плохо коррелируют с числом преступлений.

Также стоит отметить, что логарифмирование отдельных показателей не сильно улучшило ситуацию, хотя они и изменились немного в большую сторону.

Парные регрессии

Дополнительно построим парную регрессию между переменными для проверки взаимосвязи:

Vcrime ~ PR

```
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   95.991     85.818  1.1185 0.268902
PR          1899.409    597.520  3.1788 0.002589 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vcrime ~ UR

```
t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   54.690     87.597  0.6243 0.5353656
UR          5354.747   1491.795  3.5895 0.0007762 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vcrime ~ logPIPC

```
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2109.77    1421.32  1.4844   0.1442
logPIPC      -162.62     132.30 -1.2292   0.2250
```

Vcrime ~ logRGDP

```
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.0817   227.7065  0.0311   0.9753
logRGDP      29.1599    18.5975  1.5679   0.1235
```

Vcrime ~ Density

```
t test of coefficients:

              Estimate Std. Error t value Pr(>|t|)
(Intercept) 367.549568  24.542113 14.9763   <2e-16 ***
Density      -0.023391   0.073775 -0.3171   0.7526
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vcrime ~ Marriage

```
t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 304.6331    41.5103  7.3387 2.228e-09 ***
Marriage      7.9101     5.0072  1.5797    0.1207
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vcrime ~ Immigrants

```
t test of coefficients:

              Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.5287e+02 2.2127e+01 15.9472   <2e-16 ***
Immigrants  4.7359e-04 5.0370e-04  0.9402   0.3518
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vcrime ~ Temperature

```
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.0228   115.4520  0.9443  0.34974
Temperature   4.7912     2.1506  2.2278  0.03061 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vcrime ~ Divorce

```
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  128.364    114.017  1.1258  0.26583
Divorce       25.895     12.419  2.0850  0.04241 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Проверив t-test для каждой переменной отдельно, можно определить их значимость по значению p-value.

Таким образом, мы получили аналогичный предыдущему пункту результат, а именно: наиболее значимыми для анализа показателей Vcrime являются переменные PR и UR, при этом существует небольшая связь между с переменными Temperature и Divorce.

Проверка данных на гетероскедастичность

Так как графики данных неоднозначны, а наши предположения затрагивают только некоторые переменные, стоит проверить их на наличие гетероскедастичности с помощью теста Бройша-Пагана, где p-value < 0,5 свидетельствует о гетероскедастичности данных (гипотеза о гомоскедастичности отвергается).

Vcrime ~ PR

```
    studentized Breusch-Pagan test

data:  lm1
BP = 0.032846, df = 1, p-value = 0.8562
```

Vcrime ~ UR

```
    studentized Breusch-Pagan test

data:  lm2
BP = 3.2027, df = 1, p-value = 0.07352
```

Vcrime ~ logPIPC

```
    studentized Breusch-Pagan test

data:  lm3
BP = 0.21666, df = 1, p-value = 0.6416
```

Vcrime ~ logRGDP

```
    studentized Breusch-Pagan test

data:  lm4
BP = 4.5768, df = 1, p-value = 0.03241
```

Vcrime ~ Density

```
    studentized Breusch-Pagan test

data:  lm5
BP = 2.5026, df = 1, p-value = 0.1137
```

Vcrime ~ Marriage

```
    studentized Breusch-Pagan test

data:  lm6
BP = 0.053129, df = 1, p-value = 0.8177
```

Vcrime ~ Immigrants

```
    studentized Breusch-Pagan test

data:  lm7
BP = 3.2541, df = 1, p-value = 0.07125
```

Vcrime ~ Temperature

```
    studentized Breusch-Pagan test

data:  lm8
BP = 4.7504, df = 1, p-value = 0.02929
```

Vcrime ~ Divorce

```
    studentized Breusch-Pagan test

data:  lm9
BP = 2.6212, df = 1, p-value = 0.1054
```

Таким образом, показатели гетероскедастичны для переменных логарифма ВВП и температуры. Для сглаживания данного эффекта в итоговой модели будем считать робастные ошибки.

Общий вывод

Нам удалось собрать датасет с усредненными данными за обширный период времени, на основе которого мы сможем проверить нашу гипотезу о наличии зависимости между экономическими условиями штата и количеством насильственных преступлений в нем. В рамках разведанализа нам удалось провести необходимую чистку и изменение данных и убедиться в правильности выбора переменных в рамках нашего исследования.